SUPPLEMENT TO

British
Telecommunications
Engineering

ISSN 0262-4028

Vol. 6 Part 3 October 1987

THE PURPOSES OF TELETRAFFIC ENGINEERING AND ITS APPLICATION

D. J. SONGHURST, B.SC., M.SC.†

---

† Performance Engineering Division, British Telecom Engineering and Procurement.

This paper was prepared for the Institution of Electrical Engineers' (IEE) Eighth Vacation School on Switching and Signalling in Telecommunication Networks, and is published here with the permission of the IEE.

Supplement to *British Telecommunications Engineering*, Vol. 6, Part 3, October 1987.

© The Institution of Electrical Engineers.

INTRODUCTION

What is teletraffic engineering all about, why do we need it, and how do we do it? These are the questions that this paper addresses. In general, it can be said that traffic problems arise in any system that is subject to unpredictable fluctuating demands and that does not have the capacity to handle the maximum possible demand adequately. Teletraffic theory is therefore based on probability theory, and its application is essentially mathematical modelling, although computer simulation methods may be required to tackle complex models.

Although the mathematical theory underlying teletraffic engineering may be regarded as specialist expertise, the range of application is enormous. It has developed from the traditional areas of network dimensioning and optimisation and the design of multistage switching systems, through the modelling of processor-controlled exchange systems and packet-switched data networks, and into local area networks (LANs), multi-service networks, mobile radio, cable television, network management and intelligent call routing strategies.

It has sometimes been suggested that teletraffic engineering will no longer be required since equipment will become so cheap that gross overprovision will be possible. There are a number of reasons why this is a false generalisation:

- Network planners themselves show no sign of adopting this approach. In today's competitive environment, they are under even greater pressure to minimise costs.
- Many traffic problems are now associated with control software and cannot readily be solved simply by adding more hardware.
- Some systems have inherently limited or expensive resources; for example, mobile radio, cable television.

Nevertheless, the economics of telecommunications systems are changing, and the effort put into a teletraffic study should ideally be matched to the potential benefits. Many systems are now provided in large modules; this does not remove teletraffic problems, but it may require a new approach.

BASIC TRAFFIC MODELS

Traffic engineering was put on a sound mathematical footing by A. K. Erlang of the Copenhagen Telephone Company about 70 years ago. In the basic model, traffic is offered to a single group of circuits; for example, representing a route between two exchanges. The following assumptions are made:

A1 Calls arrive in a Poisson process with a constant rate (usually referred to as *random traffic*).

A2 Call holding times have a negative exponential distribution. (Essentially, calls terminate at random regardless of how long they have been in progress. This is in fact quite a good model.)

A3 The circuit group is full availability; that is, any arriving call can seize any free circuit.

A4 Calls that arrive when all circuits are busy are lost and have no effect on the system. In particular, they do not give rise to repeat attempts.

A5 The system is in statistical equilibrium; that is, it has been operating with the same call arrival rate for sufficiently long that the effects of initial conditions can be disregarded.

The traffic offered (A, measured in erlangs) is defined as the product of the call arrival rate and the mean call holding time (both expressed in the same time units). Clearly, if the circuit group is sufficiently large that no calls are lost, then A equals the mean number of calls in progress (the traffic carried).

On the basis of the assumptions stated above, Erlang derived the probability distribution of the number of busy circuits in the group. In particular, Erlang's loss formula gives the probability that all circuits are busy:

$$B_N = \frac{\dfrac{A^N}{N!}}{\displaystyle\sum_{i=0}^{N} \frac{A^i}{i!}},$$

where N is the number of circuits in the group.

When the offered traffic is random (assumption A1), the time congestion equals the call congestion; that is, the proportion of time that all circuits are busy equals the proportion of calls lost.

Assumption A2 (call holding time distribution) is not necessary for Erlang's loss formula, but it simplifies the derivation. In more complicated loss systems, and particularly in queueing systems, the call holding time distribution will affect the congestion probability.

Erlang's loss formula is usually computed by the following recursion:

$$B_N = \frac{AB_{N-1}}{N + AB_{N-1}}. \qquad (B_0 = 1)$$

For large N (>500) there are faster methods, such as:

$$\frac{1}{B_N} = 1 + \frac{N}{A} + \frac{N(N-1)}{A^2} + \cdots$$

This sum can be terminated when the terms become sufficiently small.

DIMENSIONING

Telecommunications systems and networks are generally planned by the separate dimensioning of individual component parts. Dimensioning involves determining the maximum traffic capacity for a given grade-of-service (GOS) objective. In telephone networks, GOS is normally defined in terms of the proportion of call attempts that are rejected owing to insufficient equipment (that is, congestion), although delay criteria such as delay to receive dial tone may also be used.

The allocation of GOS criteria to component parts of a network is something of an art. In principle, this should involve an economic optimisation subject to constraints on the overall GOS performance of the network. In practice, the relationship between overall GOS and the GOS objectives for individual components is difficult to quantify realistically because of the multi-dimensional variability of traffic in the network.

An important feature of traffic-carrying systems is that large traffic streams can generally be carried more efficiently than small ones. This is illustrated in Figure 1, which shows how traffic efficiency (in terms of traffic offered per circuit, for a loss probability of 1%) increases with route size. However, it follows also that large systems are more sensitive to traffic overload, as illustrated in Figure 2. In order to protect against overload, 'high-load' criteria are generally used. A route in a telephone network would typically be dimensioned to meet two criteria such as the following:

Normal load: 1% loss
20% overload: 5% loss

Lead times for provision of equipment tend to be measured in years rather than months, and so dimensioning is done on the basis of traffic forecasts. The accuracy of traffic measurement and forecasting procedures is therefore important. Traffic measurements are generally taken within busy hours on as many days as practicable. It is important to take a large measurement sample over the year so as to obtain good estimates of the mean, variability, and seasonality of traffic streams. Forecasting over periods of more than one year involves a variety of economic and other factors. Measurement and forecasting procedures are presented in some length in CCITT Recommendations E.500, E.501, E.502, and E.506.

TELEPHONE NETWORKS—ROUTING AND ANALYSIS

Large telephone networks are invariably based on hierarchical structures, but with varying degrees of non-hierarchical interconnection. The top tiers of such networks are often fully interconnected. Many networks use fixed routing, either throughout or with automatic alternative routing (AAR) restricted to part of the network. The use of fixed routing simplifies dimensioning and planning, but end-to-end GOS performance is sensitive to overloads and equipment failures on individual routes, thus requiring all routes to be dimensioned to overload criteria.

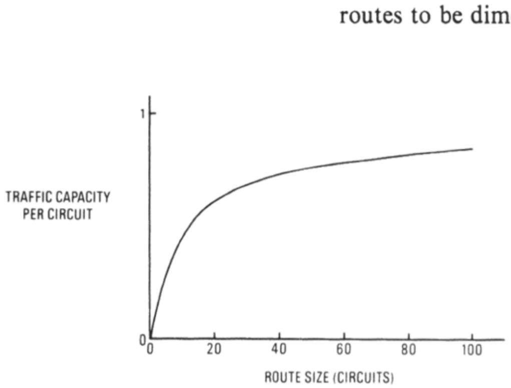


Figure 1—Traffic efficiency

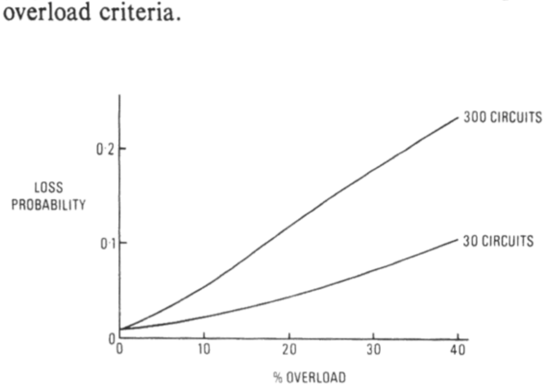
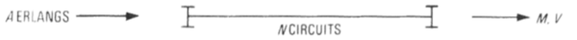


Figure 2—Overload sensitivity (1% loss at normal load)

Automatic Alternative Routing (AAR)

Many national networks use hierarchical AAR (Figure 3). Any routing scheme that uses overflow makes the mathematical dimensioning problem more difficult. Overflow traffic is non-random, and so Erlang's loss formula does not apply to routes offered overflow traffic. The dimensioning of hierarchical AAR networks can generally be tackled using decomposition and moment-matching. When random traffic is offered to a full availability route, the statistical moments of the overflow traffic (in particular, the mean M and variance V) are readily calculated as follows:

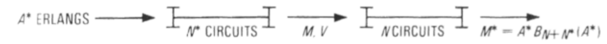


$$M = AB_N(A), \qquad V = M\left(1 - M + \frac{A}{N + 1 + M - A}\right).$$

(Warning: This variance is a theoretical parameter referring to independent observations of the traffic process. It is used for calculation purposes and is not readily observable in practice.)

Direct high-usage routes (AB and AC in Figure 3) are provided according to economic criteria. The moments of traffic overflowing from these routes are calculated from the above formulae. These can be accumulated (on the reasonable assumption of statistical independence) to give the total mean and variance of traffic offered to hierarchical routes (such as AE). Traffic that is offered directly to hierarchical routes (for example, traffic from A to D) is random and has variance = mean.

Given the statistical characteristics of traffic offered to a final-choice route, various approximate methods are available to calculate blocking and hence to dimension the route. The best-known is Wilkinson's equivalent random traffic (ERT) method [1]. A single 'equivalent' high-usage route is modelled. This is a hypothetical group of N^* circuits offered A^* erlangs of random traffic, and whose overflow moments (M, V) match those of the total traffic offered to the final-choice route.



A^* and N^* can be calculated from an iterative search, but Rapp's approximation [2] provides good estimates. The mean traffic overflowing from the final-choice route (M^*) is obtained from Erlang's loss formula, and the blocking on that route is then given by M^*/M. The route is dimensioned by choosing N to give the required blocking.

Another simple method has been developed from a suggestion by Hayward [3, 4]. If the total traffic offered to the final route has mean M and variance V, then the peakedness is defined as $z = V/M$. Random traffic has peakedness = 1 (assuming negative exponentially distributed call holding times) and overflow traffic has peakedness > 1. Final route blocking is estimated from the Erlang's loss formula for N' circuits offered A' erlangs, where $N' = N/z$ and $A' = M/z$. This approach requires extension of Erlang's loss formula to non-integer circuits, but it works surprisingly well and is adaptable to more complex situations.

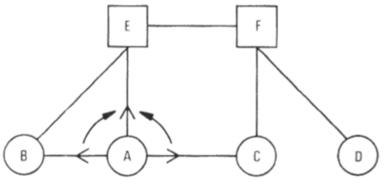


Figure 3—Hierarchical AAR

Service Protection

In networks using AAR, unrestricted overflow is rarely permitted since this can give very poor performance in overload conditions. Referring again to Figure 3, if traffic from A to B or from A to C overloads, the extra traffic almost all overflows to the hierarchical path. This has two consequences: firstly, traffic that has access to only a hierarchical route (for example, traffic from A to D) receives a very poor GOS; and secondly, calls overflowing from a direct route are using more circuits on the hierarchical path, and hence spreading the overload and potentially causing many other calls to fail. The solution to this problem is to restrict overflow traffic, for which purpose two methods, illustrated in Figure 4, are well-known.

The traditional method is to split the final route into two components, one of which forms a separate high-usage route reserved for first-choice-final traffic, the other forming a common component to which all traffic overflows. This arrangement can be analysed by using the techniques described above.

A more modern method implementable in stored-program control (SPC) systems is known as *trunk reservation*. Calls overflowing from a high-usage route to a final route are blocked whenever the final route has few free circuits (less than a specified bound). First-choice-final traffic always has full access to the final route. This approach is much more robust than the split-route method, but the analysis techniques described previously are not readily adaptable. More sophisticated techniques, such as the use of interrupted Poisson processes [5], are necessary for accurate analysis.

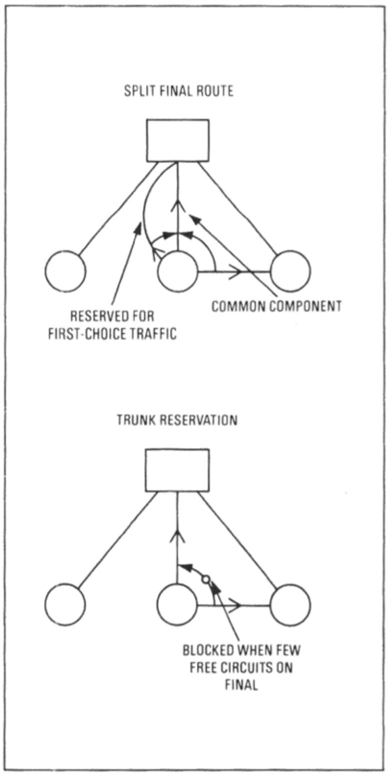

SPLIT FINAL ROUTE
RESERVED FOR FIRST-CHOICE TRAFFIC
COMMON COMPONENT
TRUNK RESERVATION
BLOCKED WHEN FEW FREE CIRCUITS ON FINAL

Figure 4—Service protection methods

Non-Hierarchical Routing

There is growing interest in the use of non-hierarchical routing methods, which include crankback (that is, returning control to the previous exchange on meeting congestion), load-sharing, dual-parenting, and dynamic routing. These can offer more flexibility in dealing with traffic overloads and equipment failures, but can be more difficult to plan and dimension. The major problem is that the network can no longer be decomposed and analysed sequentially, as with hierarchical AAR. Iterative methods using one-moment models are generally used [6, 7, 8].

Dynamic routing methods range from simple time-switched routing schedules, which form the basis of AT&T's dynamic non-hierarchical routing system (DNHR) [9, 10], to adaptive methods based on central control [11]. Recent research indicates that good results can be obtained with relatively simple distributed adaptive methods with no central control and a minimum of inter-exchange signalling [12, 13].

QUEUEING SYSTEMS

In a pure loss system, demands either receive service immediately or are lost. In a queueing system, a number of waiting places are provided where demands are queued until a server becomes free. The traditional example of a telecommunications queueing system is the operator switchboard, but the most important applications of queueing theory are now to packet-switched data networks and to the modelling of software tasks in a processor-controlled switching system. Real systems can never provide unlimited waiting places and must therefore include some form of loss mechanism (even if that is simply the failure of the system under overload conditions!). Nevertheless, pure delay models are very useful in that they can be much simpler than delay/loss models and can often adequately represent the performance of real systems in the ranges of interest. Kleinrock's two-volume treatise [14] provides a thorough survey of queueing theory and its application to data networks, although it lacks the more recent developments in queueing network theory.

Simple Queues

In the Kendall notation the simplest queue is denoted *M/M/1*, where M stands for *Markovian* (indicating the lack-of-memory property of the call arrival and termination processes that is sometimes referred to as *randomness*). The first M denotes a Poisson arrival process, the second M denotes a negative exponential service-time distribution, and the 1 refers to a single server. It is also assumed that the number of waiting places is unlimited and that service is given in order of arrival (first-come first-served). It is then found that, in statistical equilibrium, the number of demands in the system has a geometric distribution, and the delay time has a negative exponential distribution with a discrete peak at zero (corresponding to demands that are served immediately).

$$\text{Mean number in system} = \frac{A}{1-A}.$$

$$\text{Server occupancy} = A.$$

$$\text{Probability of delay} = A.$$

$$\text{Mean delay time} = \frac{AH}{1-A}.$$

In the above equations, H is the mean service time, and A is the mean arrival rate of demands per mean service time (= offered traffic).

Now a pure delay system cannot attain statistical equilibrium (that is, it is unstable) unless the offered traffic is less than the number of servers. For the single-server queue, this requires $A < 1$.

Relatively simple explicit results for delay distribution are also available for the first-come first-served multi-server queue M/M/N with limited or unlimited waiting places. One of the few other simple results in queueing theory is the Pollaczek-Khinchine equation, which applies to the M/G/1 queue; that is, a single-server queue with a general service-time distribution:

$$\text{mean delay time} \quad = \frac{A(1+C)H}{2(1-A)},$$

where C is the coefficient of variation of the service time distribution; that is, the variance divided by the square of the mean. For the M/M/1 queue $C = 1$. For the M/D/1 queue, with deterministic (that is, constant) service times, $C = 0$, and therefore the mean delay is precisely half that of the M/M/1 queue.

The performance of some simple queues is illustrated in Figure 5. This shows firstly that the 10-server queue M/M/10 is more efficient than the single-server queue M/M/1, and correspondingly more sensitive to overload. Also shown is the 10-server queue with only 10 queue places, M/M/10/10. In this queueing system, calls that arrive to find all queue places occupied are rejected, and the mean delay for accepted calls is thereby held to a finite limit.

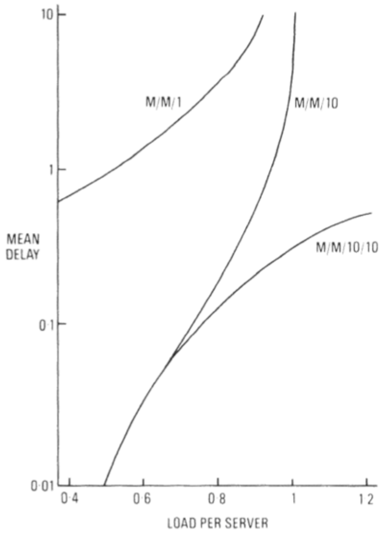


Figure 5—Performance of some simple queues

Networks of Queues

Many systems of practical interest are effectively networks of queues, where demands may move from one queue to another, receiving service at each. Although such systems can be exceedingly complex to analyse, a number of remarkably simple results have appeared in recent years. The starting point is the fact that the output (that is, the departure process) from the multi-server M/M/N queue is itself a Poisson process. Hence, if demands leave this queue after being served and join a second queue with exponential service times, then the second queue is itself M/M/N. We can use this result to calculate the mean sojourn time (that is, the time that any demand spends in the system, including all queueing times and service times) for a queueing network under the following assumptions:

A1 All queues are first-come first-served with exponentially-distributed service times and unlimited waiting room.

A2 Fresh demands may arrive at any queue in a Poisson process with constant rate.

A3 Demands move between queues according to a deterministic or probabilistic (but not state-dependent) routing rule.

A4 There is no feedback; that is, the queues can be strictly ordered.

Jackson [15] showed that assumption A4 is not actually necessary. When feedback is allowed, demands may return to a queue where they have previously been served. Although this destroys the Poisson property of the output process, each queue in the network still behaves as if it were M/M/N. Baskett *et al* [16] have extended this to even more general networks (now known as *BCMP* networks, after the initials of the authors) by allowing certain relaxations in assumption A1.

In the analysis of queueing networks, a simple and general result known as *Little's formula* is particularly useful. This states that, for any system in equilibrium:

$$N = LT,$$

where N = mean number of demands in the system,
L = mean arrival rate, and
T = mean sojourn time of a demand.

This can be applied to a queueing network as a whole, and to any part of the network individually. Reiser [17] has used this to develop a method of mean value analysis which applies to closed BCMP networks (closed in the sense that no fresh demands arise—there is simply a fixed number of demands circulating in the network), and is useful as an approximation for more general networks.

Packet-Switched Data Networks

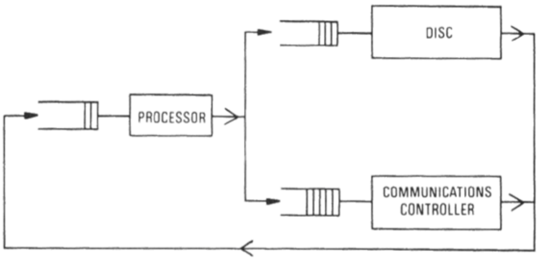
Many countries now have public data networks based on packet switching. Such networks operate on the store-and-forward principle. Each packet-switching exchange has buffer space in which packets are stored while awaiting or receiving processing, while queueing for onward transmission, and while awaiting acknowledgement from the next exchange. A packet-switched network is therefore a complex queueing network, whose primary resources are buffer space, processor time, and link bandwidth. Queueing theory is widely used for the analysis of data networks, to tackle problems such as the following:

(*a*) optimisation of buffer space allocation,
(*b*) analysis of processing delays,
(*c*) analysis of protocols such as X.25 which determine the effective link transmission times dependent on error rates and acknowledgment procedures,
(*d*) analysis of end-to-end packet transmission times,
(*e*) design of network structures and routing procedures, and
(*f*) analysis of flow control procedures which protect the network from overload by limiting the rate at which packets or new calls are accepted.

Processor-System Modelling

When setting up a call, an SPC system uses a number of software processes. These run on processors (possibly only one) and interact with storage devices and various input/output media such as inter-exchange signalling systems. Queueing models can often be used to analyse problems in such systems, such as store-contention. A simple queuing network model for a processor system is shown in Figure 6. Queueing network models can generally be applied to provide estimates of processor occupancy and mean throughput time. However, many such systems are complex multi-processor systems with multi-priority processes. Moreover, proper dimensioning requires analysis of delay distributions, for which computer simulation methods are often necessary.

Figure 6
Simple processor system queueing model



SWITCHING SYSTEM DIMENSIONING

Performance requirements for a processor-controlled exchange are specified in terms of loss probabilities and delay times. Losses may be due to switch blocking or to the processor control. Switch blocking is a significant factor in the concentration stages of local exchanges, whereas large digital tandem switches tend to be designed as virtually non-blocking.

Various delay criteria are specified, including dial tone delay (in local exchanges), incoming response delay, and switching delay. These delays are a function of processor performance, and since processors act as queueing systems, they are sensitive to overloads. Well-designed processor-controlled systems have a load-control mechanism that rejects call attempts before delays become unmanageable. The processor load is determined by the call attempt rate and the switch load is determined by the offered traffic; therefore, the mean call holding time is an important factor in exchange dimensioning. In overload conditions, more calls are ineffective and generate repeat attempts, so it is usual to specify higher overload levels for call attempts than for offered traffics (typically 40% and 20% respectively).

There are therefore a number of requirements for switching system dimensioning:

- Performance requirements must be specified for normal load, overload, and partial failure conditions.
- A good model for exchange behaviour must be available.
- Information is required on mean call holding time, call mix (that is, the relative proportions of different types of calls), and the characteristics of traffic overloads.

NEW APPLICATIONS

Teletraffic engineering is undergoing rapid development for application to a range of new kinds of telecommunication systems, such as the following:

Integrated Services Digital Networks

The developing integrated services digital network (ISDN) will bring a whole range of new traffic problems, many of which cannot yet be foreseen in detail. Areas that have already received substantial study include the performance of access protocols and the dimensioning of circuit groups carrying traffics with multiple bandwidths [18].

Local Area Networks

LANs are either based on bus designs (for example, Ethernet) or ring designs (for example, Cambridge and Orwell Rings [19]). Random-access protocols, such as that used in Ethernet, have interesting traffic properties and are used in a variety of other applications such as cellular mobile radio signalling systems [20]. Ring designs are basically cyclic queueing systems which tend to be very complex to analyse. LANs are generally designed to operate at relatively low occupancies.

Cellular Mobile Radio

In addition to the problems of random-access signalling systems referred to above, the optimal allocation of radio channels to cells can be a difficult problem. Call handover from one cell to another is a form of overflow traffic whose characteristics depend on the mechanism for initiating it and where blocking is not allowable. The analysis of advanced systems using dynamic channel allocation involves particularly complex topological considerations.

TOOLS FOR TELETRAFFIC ANALYSIS

Mathematical methods are available for a wide range of teletraffic problems. Although many of these methods may sometimes be regarded as 'exact', there are a number of reasons why they will never predict precisely what happens in practice:

- The models are statistical and give mean values that correspond to indefinitely long periods of observation. Congestion performance tends to be highly variable within limited periods (for example, 1 hour), although this variability can itself be predicted from mathematical models.

- Real traffic behaviour is unpredictable and may not match the assumptions of the model. For example, the effects of day-to-day traffic variability are often not included in dimensioning models.
- It is often necessary to use simplified models of system behaviour to enable mathematical analysis.

This last problem can be attacked by using computer simulation. The accuracy of the system model is then limited only by the information available from system designers, the effort available for program development, and the number of bugs in the program. It is important to remember that, even with an accurate system model, a simulation is still an attempt to solve a mathematical problem; that is, how does the system perform under a certain mathematical model of traffic behaviour. Furthermore, simulation is an inherently statistical method, and long computer runs are often required to give results with sufficient accuracy. Nevertheless, simulation is virtually essential for detailed analysis and dimensioning of all but very simple systems.

Ideally, simulation should always be supplemented by simplified analytical models. These can usually be programmed and run far more quickly and cheaply so as to give a better understanding of how the system performance depends on different factors, and to help uncover bugs in the simulation. Analytical models can be calibrated by simulation, and then used for system dimensioning after the design and evaluation phases are complete. A number of packages are available commercially, notably QNAP (*Q*ueueing *N*etwork *A*nalysis *P*ackage), which provides analytical and simulation methods for queueing systems.

References

1 WILKINSON, R. I. Theories for toll traffic engineering in the USA. *Bell Syst. Tech. J.*, 1956, **35**, pp. 421–514.
2 RAPP, Y. Planning a junction network in a multi-exchange area. *Ericsson Technics*, 1964, **20**(1), pp. 77–130.
3 FREDERICKS, A. A. Congestion in blocking systems—a simple approximation technique. *Bell Syst. Tech. J*, 1980, **59**, pp. 805–827.
4 LINDBERGER, K. Simple approximations of overflow system quantities for additional demands in the optimisation. ITC 10, paper 5.3.3 1983.
5 SONGHURST, D. J. Protection against traffic overload in hierarchical networks employing alternative routing. Proc. Telecommunication Networks Planning Symposium, Paris, 1980, pp. 214–220.
6 GIRARD, A., and COTE, Y. Sequential routing for circuit-switched networks. *IEEE Trans. Commun.*, 1984, **32**(12), pp. 1234–1239.
7 LEGALL, F., and BERNUSSOU, J. An analytical formulation for grade of service determination in telephone networks. *IEEE Trans. Commun.*, 1983, **31**(3), pp. 420–4.
8 LEBOURGES, M., BECQUE, C. R., and SONGHURST, D. J. Analysis and dimensioning of non-hierarchical telephone networks. ITC 11, paper 2.2B-4, 1985.
9 ASH, G. R., CARDWELL, R. H., and MURRAY, R. P. Design and optimisation of networks with dynamic routing. *Bell Syst. Tech. J.*, 1981, **60**(8), pp. 1787–1820.
10 ASH, G. R., KAFKER, A. H., and KRISHNAN, K. R. Intercity dynamic routing architecture and feasibility. ITC 10, paper 3.2.2 1983.
11 CAMERON, W. H., *et al.* Dynamic routing for intercity telephone networks. ITC 10, paper 3.2.3, 1983.
12 GIBBENS, R. J. Aspects of dynamic routing in circuit-switched telecommunication networks. Third UK Teletraffic Symposium, University of Essex, June 1986.
13 STACEY, R. R., and SONGHURST, D. J. Dynamic alternative routing in the British Telecom trunk network. International Switching Symposium, Phoenix, 1987.
14 KLEINROCK, L. Queueing Systems. Vol I: Theory, Vol II: Computer Applications. J. Wiley, 1975.
15 JACKSON, J. R. Jobshop-like queueing systems. *Manage. Sci.*, 1963, **10**, pp. 131–142.
16 BASKETT, CHANDY, MUNTZ, and PALACIOS. Open, closed and mixed networks of queues with different classes of customers. *J. ACM.*, 1975, **22**, pp.248–260.
17 REISER, M., and LAVENBERG, S. S. Mean value analysis of closed multichain queueing networks. *ibid.*, 1980 **27**, pp. 312–322.
18 JOHNSON, S. A. A performance analysis of integrated communications systems. *Br. Telecom Technol. J.*, Oct. 1985, **3**(4), pp. 36–45
19 FALCONER, R. M., and ADAMS, J. L. Orwell: a protocol for an integrated services local network. *ibid.*, October 1985, **3**(4), pp. 27–35.
20 MACFADYEN, N. W., and EVERITT, D. E. Teletraffic problems in cellular mobile radio systems. ITC 11, paper 2.4B-1, 1985.

Bibliography

A good general introduction to teletraffic engineering is given by D. Bear, in *Principles of telecommunication-traffic engineering*, IEE Telecommunications Series 2, Peter Peregrinus (1976).

The classic text book is R. Syski's *Introduction to congestion theory in telephone systems*, Oliver and Boyd (1960, revised edition 1986).

The main forum for the publication of teletraffic work is the International Teletraffic Congress (ITC). Copies of the proceedings are available at the Science Reference Library. The most recent (11th) congress proceedings are published by North Holland. The UK Teletraffic Symposium is a developing national forum which is linked to the ITC and seeks to promote interactions in the UK.